Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 8, pp. 2227–2236
DOI : 10.47974/JIOS-1785

An enhanced Grey Wolf optimizer Cuckoo search optimization with Naïve Bayes classifier for intrusion detection system

Satish Kumar Garapati *
A. N. Sigappi †
Department of Computer Science and Engineering
Annamalai University
Annamalainagar 608002
Tamil Nadu
India

Abstract

Firewalls, cryptographic methods, and antivirus scanners are becoming increasingly ineffective in the face of increasingly complex threats. To make network channels and machines more secure, stronger protective walls are required. In addition to the policies already in place, an intrusion detection system can serve as an extra line of defence. This research suggests using a Naïve Bayes classifier in conjunction with an Enhanced Grey Wolf Optimizer Cuckoo Search Optimization (EGWCSO) to enhance performance. Min-Max normalization and 1-N encoding are used in the preprocessing step. EGWCSO is applied for the best feature selection, and Naïve Bayes is selected for classification. Single connection vectors, around 4,900,000 are present in the NSL KDD dataset. The performance of the classifier is checked against them. It consists of 41 attributes, 5 normal classes, and 4 attack types namely Dos, Probe, R2L, and U2R. The experimental result concludes the proposed EGWCSO with Naïve Bayes classifier gives higher precision, recall and accuracy values and less execution time compared to previous algorithms.

Subject Classification: *68T07, 94A60.*

Keywords: *Cuckoo search optimization, Grey wolf optimizer, Naïve bayes classifier, NSL KDD dataset.*

1. Introduction

With the increase in online banking, e-commerce, and telecommunication, internet has become a powerful platform for the

* *E-mail:* **gskchowdari@gmail.com (Corrsponding Author)**
† *E-mail:* **an.sigappi@gmail.com**

dissemination of knowledge, commerce, entertainment, and information. However, cyber-attacks are soon turning out to be a pressing global concern as they pose an issue to data security across both personal and business fronts [1]. Cybercrime is on the rise due to e-commerce in India [2]. For the mitigation of risks, firewalls, authentication mechanisms, rights of access, encryption, and effective IDS can be used to safeguard computers. IDS watches user activity for suspicious behavior, sounds alarms, defines alarm-generating events, and notifies the administrators about security incidents.

2. Related work

Sangve et al. [3] have for research purposes, the NSL-KDD dataset on the utilization of techniques of Machine Learning in Anomaly Network Intrusion Detection Systems (ANIDS). Zhu et al. [4] proposed the U-D methodology for upgrading intrusion detection through uploading and downloading data, and the method has demonstrated a quite satisfactory efficiency of detection. Aziz et al. [5] applied the genetic algorithm along with different strategies relative to feature selection and found out that the sequential floating algorithm proved to be the most effective, with detection accuracies of 92.86% and 85.38%. These authors, Seth et al., applied in Ref. [6] Binary Grey Wolf Optimization and classifier of a neural network to reduce the number of dataset instances and increase accuracy by removing unimportant features. An IDS based on deep learning combined with DNN and deep autoencoders PTDAE with hyperparameter optimization was proposed by Kunang et al. in Ref. [7]. R. Sekhar et al. [8] proposed a method for intrusion detection through Deep Autoencoders with Fruitfly Optimization and the Fuzzy C-Means Rough Parameter technique, evaluated over the NSL-KDD and UNSW-NB15 datasets.

3. Proposed System

In the proposed system, the GWO-CS with Naïve Bayes classifier is introduced to obtain more accurate classification result for NSL KDD cup dataset. The proposed system architecture is illustrated in Figure 1.

3.1 *Dataset*

The NSL-KDD dataset is suitable for comparative analysis with 41 features and having 5 classes, with 4 types of attacks: DoS, Probe, R2L, U2R, and the benefits of duplication elimination, compatible with many

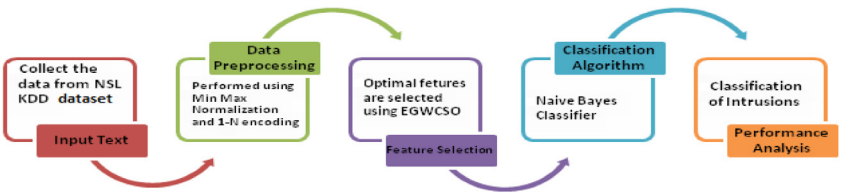

Figure 1

Proposed Architecture

learning algorithms, and also enough instance numbers of both training and testing data sets for economical experimentation.

3.2 *Preprocessing*

The NSL-KDD dataset undergoes data pre-processing utilizing 1-N encoding together with min-max normalization for data normalization and classification, respectively.

3.2.1 Min-max normalization

The normalization process minimizes feature scales and dimensions, modifies data by a small range through linear transformation [9], and normalizes dimension values up to [0, 1] using min-max normalization, as described in Eqn. (1).

$$m(ia) = \frac{d - \min_k}{\max_k - \min_k}(T\min_k - T\max_k) + T\min_k \quad (1)$$

Let, d represent the data value's converted rate and k can be taken as a measurement or characteristic. In addition, k is indicated as the initial lowest amount with k and max$k$ is indicated as the initial maximum amount with k. Additionally, Tmin_k is represented as a modified minimum value using k and Tmax_k is taken as a transformed maximum value with k.

3.2.2 1-N encoding

1-N encoding is an important step in transforming variables for categorical data into DL, which improves the precision of identifying and categorizing. It converts variables into numerical values [10], with each potential category having its binary features. Each sample feature matches the original category, assigning a 1 value. The feature set can range from 0 to n-class 1.

3.3 *Feature selection*

3.3.1 Grey Wolf Optimization

The Grey Wolf Optimizer (GWO) is a meta-heuristic algorithm [11] that mimics grey wolf hunting techniques through three steps: encircling prey, hunting, and attacking. It follows the social hierarchy of alphas, betas, and omegas, with deltas representing seniors, searchers, and guardians [12]. Equations 2 and 3 simulate wolves' surrounding behaviour during the hunt.

$$\vec{D} = \vec{B}.\overrightarrow{P_{prey}}\ (i) - \overrightarrow{P_{wolf}}\ (i) \tag{2}$$

$$\overrightarrow{P_{wolf}}(i+1) = \overrightarrow{P_{prey}}(i) - \vec{A}.\vec{D} \tag{3}$$

Where i represents the current iteration, and $\vec{A}$ and $\vec{B}$ represents coefficient vectors, $\vec{P}$ refers to the position vector of the prey. The vectors $\vec{A}$ and $\vec{B}$ are computed as in Equation (4) and (5).

$$\vec{A} = 2\vec{a}.\overrightarrow{r1} - \vec{a} \tag{4}$$

$$\vec{B} = 2\overrightarrow{r2} \tag{5}$$

The Grey Wolf Optimizer (GWO) optimizes the hunt by reducing normal vectors in the interval [0,1]. Alpha leads the hunt, beta and delta occasionally join. The GWO saves alpha, beta, and delta's best solutions, while omega agents update positions [13].

3.3.2 EGWCSO algorithm

The CS algorithm enhances GWO dependability and search execution by accelerating particles as they converge on the best solution and using random disposal to reduce local optima for more efficient search.

$$p^{i+1} = p(i) \oplus Levy(\beta) \tag{6}$$

The step size, denoted by $\beta>0$, is determined by the issue scales and can be used in various situations. In multiplication, the product shows an entry-wise walk, while when Levy Flight random steps are limited, the walk is typically random [14].

Pseudo code for EGWCSO algorithm

GWO is a quality reduction technique that uses efficient filter-based target-hunting, principles that make use of the classification results of wrapper-based techniques [15]. This GWO is bi-stage: To optimize the exchange of data, GWO seeks out combinations of attributes.

$$\theta = V - P \tag{7}$$

The GWO classifier's accuracy is evaluated using mutual information (P), with convergence determined by search strategy and fitness function, and parameter 'a' regulating exploration and exploitation.

$$\text{Fitness} = \text{CCR (D)} \tag{8}$$

The Naïve Bayes classifier enhances classification accuracy by calculating prior and conditional probabilities from the training dataset D, sorting samples based on posterior probabilities, and classifying instances into DoS, R2L, U2R, and PROBE attack types [16].

3.3.3 Naïve Bayes Classifier

A supervised machine-learning algorithm is the naive Bayes classifier algorithm technique that uses probabilistic computations to predict binary outputs from a data set, using a frequency table to compute each class's posterior probability.

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) * P(A)}{P(B)} \tag{9}$$

The optimal classifier is determined using a dataset, and the Naive Bayes model generates an accurate model. The suggested system enhances performance metrics like recall, accuracy, precision, and f-measure [17] [18].

4. Experimental Result

4.1 *Accuracy*

Accuracy refers to the accuracy of a quantity's calculation, measurement, or specification, and the degree to which the model accurately demonstrates the intrusion detection system's efficacy by being close to the genuine value. Accuracy can be measured by comparing measurement results with true value results using Eqn. 10.

$$A = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{10}$$

Let, T_p be denoted as the true positive of intruders' activity of correctly classified attacks, T_n is considered as the true negative of classifying normal activity as positive, F_p is represented as false positive of incorrect classification of normal activity as an attack, and F_n is denoted as false negatives of classifying the intrusive activity as normal activity. Figure 2 shows the comparison of accuracy.

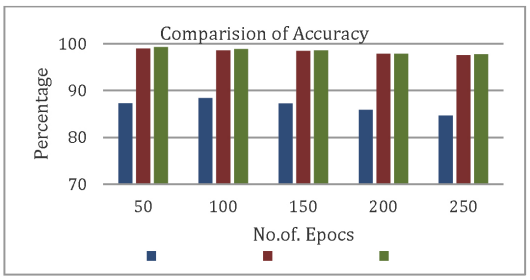

Figure 2

Comparison of Accuracy

4.2 *Precision*

Precision is typically defined as how much a tool or development will repeat a given value. The amount of accurate intrusion predictions that are in line with overall positive projections is what this method is measured for. Additionally, the proportion of intrusion detection, or precision, is calculated using Equation (11). Figure 3 shows the comparison of precision.

$$P = \frac{T_p}{T_p + F_p} \tag{11}$$

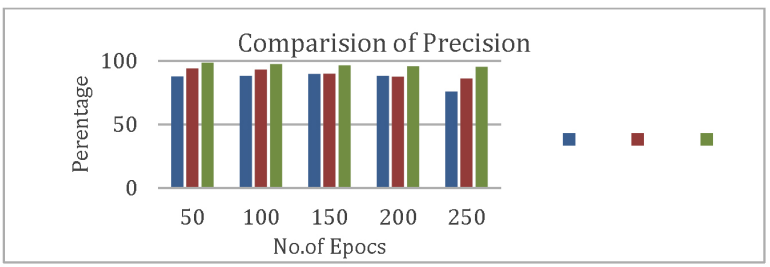

Figure 3

Comparison of Precision

4.3 *Recall*

The capacity to accurately identify and detect network intrusions is known as recall. The number of true positives that are accurately predicted is called recall. Figure 4 demonstrates the comparison of recall.

$$S_N = \frac{T_p}{T_p + F_n} \tag{12}$$

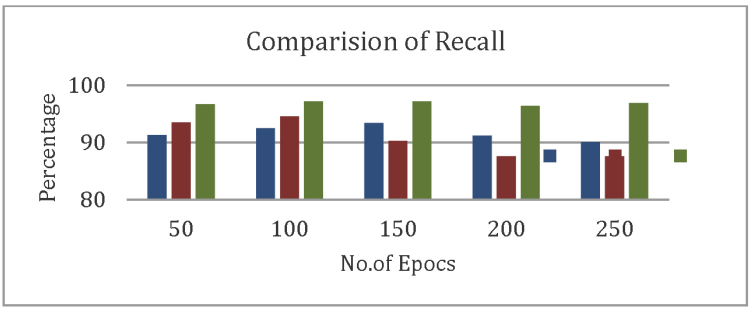

Figure 4

Comparison of Recall

4.4 *FPR*

FPR or IDS interprets typical activity as an attack or intrusion, whereas a false alarm is a negative statement signaling a network incursion. By forecasting the likelihood of erroneously anticipating the negative class, it calculates accuracy. Figure 5 and Figure 6 shows the comparison of FPR and execution time of various methods.

$$FPR = \frac{F_n}{T_p + F_n} \tag{13}$$

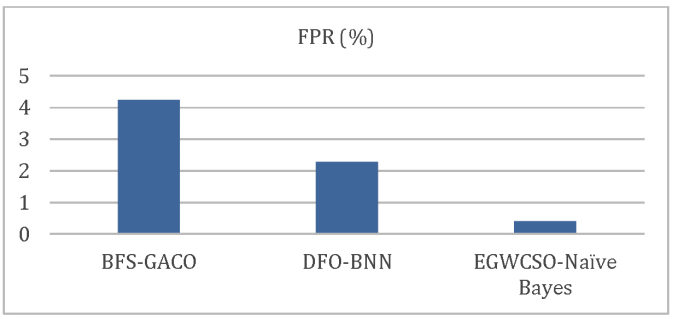

Figure 5

Comparison of FPR

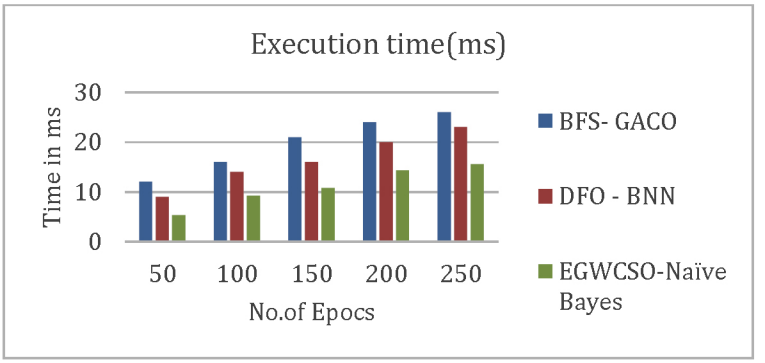

Figure 6

Comparison of execution time.

5. Conclusion

This study enhances intrusion detection accuracy by integrating EGWCSO with the Naïve Bayes algorithm. It addresses previous shortcomings in attack detection by introducing improved optimization and classification techniques. The approach includes feature selection using EGWCSO on the dataset from the NSL KDD, then min-max normalization for preprocessing. Selected features are refined to enhance classification performance using Naïve Bayes. Evaluation metrics such as precision, recall, and accuracy demonstrate improved results with EGWCSO. Experimental findings indicate superior performance compared to existing methods, suggesting potential for more effective attack discovery in future systems combining advanced optimization and classification algorithms.

References

[1] Y. Zhao, "Network intrusion detection system model based on data mining," paper presented at the 7th IEEE/ACIS Int. Conf. Software Eng., Artif. Intell., Networking and Parallel/Distrib. Comput. (SNPD), Shanghai, China (2016).

[2] V. Kumar, J. Srivastava, and A. Lazarevic, Eds., Managing Cyber Threats: Issues, Approaches, and Challenges, vol. 5. Springer Science & Business Media (2006).

[3] S. M. Sangve and R. Thool, "A formal assessment of anomaly network intrusion detection methods and techniques using various datasets," *IEEE Int. Conf. Appl. Theor. Comput. Commun. Technol.* (iCATccT), pp. 267–272 (2015).

[4] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access,* vol. 5, pp. 21954–21961 (2017).

[5] R. Aziz, C. K. Verma, and N. Srivastava, "A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data," *Genomics Data*, vol. 8, no. 1, pp. 4–15 (2016).

[6] J. K. Seth and S. Chandra, "Intrusion detection based on key feature selection using binary GWO," in IEEE Int. Conf. Computing for Sustainable Global Development (INDIACom), pp. 3735–3740 (2016).

[7] A. Sharma and U. Ghose, "Deep learning based bi-polar sentiment classification of movie reviews in Hindi," *J. Stat. Manag. Syst.,* vol. 27, no. 1, pp. 59–86 (2024).

[8] S. K. Mahapatra, B. K. Pattanayak, and B. Pati, "Attendance monitoring of masked faces using ResNext-101," *J. Stat. Manag. Syst*., vol. 26, no. 1, pp. 117–131 (2023).

[9] P. Hujare, P. Rathod, D. Kamble, A. Jomde, and S. Wankhede, "Predictive analytics of disc brake deformation using machine learning," *J. Inf. Optim. Sci.,* vol. 45, no. 4, pp. 1153–1163 (2024).

[10] M. Ramakrishna, A. R. Satish, and G. Ashok, "Malware strategies and issues in examination," *J. Inf. Optim. Sci.,* vol. 45, no. 4, pp. 1117–1127 (2024).

[11] A. Gupta, M. C. Lohani, and M. Manchanda, "Utilizing mathematical concepts of heat map for an intelligent and secure approach to efficiently detect credit card fraud," *J. Interdiscip. Math.*, vol. 26, no. 8, pp. 1837–1854 (2023).

[12] K. Kaur, S. Chakraborty, and A. K. Sagar, "Unleashing a cascade of machine learning for turbocharged sequence alignment in discrete mathematical sciences using GPU," *J. Discrete Math. Sci. Cryptogr.,* vol. 27, no. 4, pp. 1129–1138 (2024).

[13] J. Chandwani, G. Dhopavkar, M. Tatiya, N. Chakole, S. V. Kulkarni, and N. Shelke, "Security-aware analytical framework: A mathematical model and machine learning for dynamical system control in se-

cure environments," *J. Discrete Math. Sci. Cryptogr.*, vol. 27, no. 2-B, pp. 715–727 (2024).

[14] P. Liu and S. Zhang, "A novel cuckoo search algorithm and its application," *Open J. Appl. Sci.*, vol. 11, pp. 1071–1081 (2021).

[15] E. M. Roopa, S. R. M. Prasanna, N. J. R. Mahendra, and S. A. R. M. Latha, "Enhanced transductive support vector machine classification with grey wolf optimizer cuckoo search optimization for intrusion detection system," *Concurrency Comput. Pract. Exp.*, vol. 32 (2018).

[16] G. Li, T. Wang, Q. Chen, P. Shao, N. Xiong, and A. Vasilakos, "A survey on particle swarm optimization for association rule mining," *Electronics*, vol. 11, no. 19, art. 3044 (2022).

[17] H. Chen, S. Hu, R. Hua, Y. Zhang, Y. Zhang, and S. Wang, "Improved naive Bayes classification algorithm for traffic risk management," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 30 (2021). [Online]. Available: https://doi.org/10.1186/s13634-021-00742-6.

[18] W. Zhang and F. Gao, "An improvement to Naive Bayes for text classification," Procedia Eng., vol. 15, pp. 2160–2164 (2011). [Online]. Available: https://doi.org/10.1016/j.proeng.2011.08.404.

Received February, 2024